一种改进的深度残差网络行人检测方法 *

郝旭政，柴争义

(天津工业大学 计算机科学与软件学院，天津 300387)

摘 要：为了提高行人检测方法的准确率，针对行人图像特征，提出一种基于深度残差网络和 YOLO(You Only Look Once)方法的行人检测方法。以加强行人特征表达为目的，通过分析行人在图像中的表达和分布特征，提出一种不影响实时性的矩形输入深度残差网络分类模型以改进 YOLO 检测方法，使模型能够更好的表征行人；为了进一步提高模型的准确率和泛化能力，采用了混合行人数据集训练的方式，提取 VOC 数据集的行人数据与 INRIA 数据集组成混合数据集进行训练，明显降低了漏检率；并且利用聚类分析预测框的方法重新设计了初始预测框，提高行人定位能力并加快收敛。经公开的 INRIA 数据集的测试实验证明，本方法较主流的行人检测方法每张图片误检率有明显改善，降低至 13.86%，有 1.51%至 58.62%不同程度的提升，并且本方法拥有良好的实时性和泛化能力，实用性强。

关键词：行人识别；深度残差网络；YOLOv2；卷积神经网络；深度学习

中图分类号：TP391.41　　doi: 10.3969/j.issn.1001-3695.2017.12.0836

Improved pedestrian detection method based on depth residual network

Hao Xuzheng, Chai Zhengyi

(*School of Computer Science & Software Engineering*, *Tianjin Polytechnic University*, *Tianjin* 300387, *China*)

Abstract: To improve the accuracy of the pedestrian detection method, a rectangular input of convolution neural network enhance the new pedestrian detection method based on the depth residual network and YOLO object detection method. The rectangular input helped the model gain the pedestrian characteristics expression by analyzing the expression and distribution characteristics of pedestrians in the images. The depth residual network with pre-activation for YOLO object detection improved the feature extraction ability through more layers of convolution neural networks. Hybrid dataset training and cluster anchor boxes could also improve the pedestrian detection performance. The test results of INRIA dataset have proved that the method has better detection performance than the popular pedestrian detection methods, the index of False Positive per Image can reduce to 13.86%, improving ranging from 1.51% to 58.62% in varying degrees.

Key words: pedestrian detection; deep residual network; YOLOv2; convolutional neural network; deep learning

0 引言

行人检测是计算机视觉领域的重要研究课题之一。行人检测通常通过图像处理、计算机视觉相关算法以及机器学习等技术对行人进行检测和识别[1]。行人检测和识别技术在无人驾驶汽车、智能视觉监控、服务型智能机器人等领域中均有重要应用。

根据分类方法，行人检测通常被分为形变部件模型变体、基于决策森林的行人检测方法和基于深度神经网络的行人检测方法三种[2]。近年来，行人检测在这三个研究方向均有不同程度的发展，在公开的行人数据集上可以取得相似的性能表现。但类似方向梯度直方图(HOG)[3]、局部二值模式(LPB)[4]等手工特征的传统行人检测方法存在特征维度高、手工设计特征、泛化能力差等缺点，在实际应用中效果较差[5]。基于深度神经网络的行人检测方法能够自动学习行人特征，具有较好的泛化能力，随着卷积神经网络在目标检测领域的成功应用，引起了研究者的广泛关注，成为计算机视觉领域的研究热点之一[6]。

基于卷积神经网络的目标检测方法主要有区域提名和端到端两个方向。基于区域提名方向的研究主要是 R-CNN(region-based convolutional neural networks，基于区域的卷积神经网络)[7~9]、SPP-net (spatial pyramid pooling networks，金字塔池化层网络)[10]和 R-FCN(Region-based Fully Convolutional Network，基于区域的全卷积网络)[11]。Faster R-CNN[9]使用了区域提名网络(region proposal networks，RPN)，提出预测框(anchor box)机制。端到端方向的研究主要有 YOLO[12,13]和 SSD(single shot multi-box detector)[14]。YOLOv2[13]方法去掉 YOLO 全链接层引

收稿日期：2017-12-22；**修回日期：**2018-02-08　**基金项目：**国家自然科学基金资助项目（U1504613）

作者简介：郝旭政（1989-），男，山东昌邑人，硕士研究生，主要研究方向为深度学习，智能优化算法（hao_xuzheng@163.com）；柴争义（1976-），男，教授，博士，主要研究方向为智能优化算法，认知无线网络，物联网.

入 Faster R-CNN 提出的区域提名网络，加入聚类分析，多尺度训练、多层次特征融合层等改进，并在此基础上加入词汇向量树和联合训练算法，提出可以检测超过 9000 种物体 YOLO9000[13]方法。

本文基于端到端目标检测的研究成果，针对行人特征提出了一种融合深度残差网络[15,16]和 YOLO 方法的行人检测和识别方法。本方法通过分析行人在图像中的表达和分布特征，提出一种 9:19 矩形输入的卷积神经网络分类模型，加强了行人特征表达；为了提高模型的特征提取能力，本文提出了基于 50 层的预激活深度残差网络的 YOLO 行人检测方法，使模型能够更好的表征行人；为了提高模型的准确率和泛化能力，本文提出了混合行人数据集训练方式，手工提取了 VOC 数据集的行人数据与 INRIA 数据集组成混合数据集进行训练，明显降低网络漏检率；最后，本文针对提出的网络模型和增强的数据集通过聚类分析方法设计了初选预测框，提高行人定位能力，加快收敛。实验证明，本算法拥有良好的实时性和泛化能力，在公开的行人数据集下，本方法较主流的行人检测方法每张图片误检率有明显提升，并且本方法拥有良好的实时性和泛化能力。

1　相关知识

1.1　YOLOv2 目标检测方法

YOLOv2 目标检测方法是目前最先进的目标检测方法之一，在检测速度和检测精度均有出色表现。YOLOv2 方法使用的是基于 GoogLeNet 的定制网络，结构如图 1。其分类网络是 Darknet–19[13]网络模型，拥有 19 个卷积层和 4 个池化层，把 1×1 的卷积核置于 3×3 的卷积核之间来压缩特征，每次池化后将特征图维度加倍。分类网络还采用了批规范化、预训练和多尺度训练等方法提高网络分类准确率。

YOLOv2 的检测网络借鉴了 Faster R-CNN，使用了带有 Anchor 预测框的卷积层，并优化了先验预测框的选取；YOLOv2 还设计了多层次特征融合层，把 26×26×512 的浅层特征图拆分并组成 13×13×2048 的特征图，并连接到深层组成 13×13×3072 特征图，从而使 YOLOv2 能够获得更好的细粒度特征。

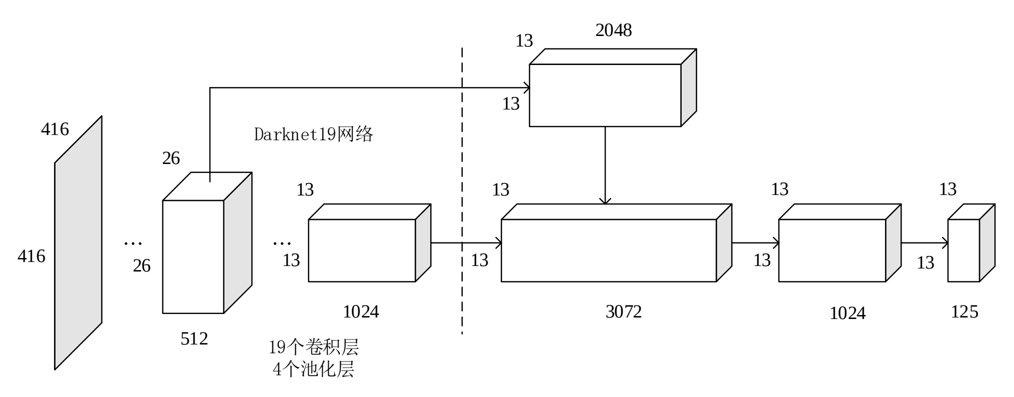


图 1　20 分类的 YOLOv2 结构图

1.2　深度残差网络

深度残差网络（Deep Residual Network, Resnet）是何恺明等人在文献[15]中提出的极深卷积神经网络模型，可以避免简单堆叠的卷积神经网络梯度消失或爆炸以及精度退化问题，模型更容易优化，性能提升明显[15]。

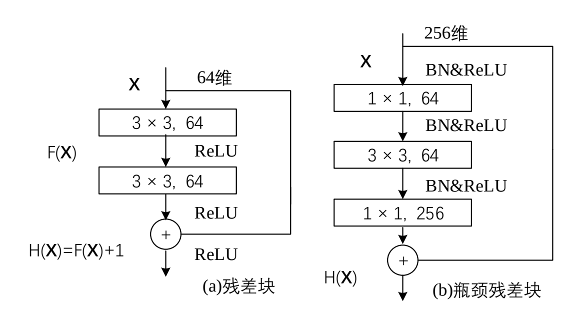


图 2　深度残差网络基本结构

假设 $H(X)=X$ 来表示最优解映射，一般的卷积神经网络是直接拟合 $H(X)=X$，而深度残差网络期望网络拟合残差映射，即

$$F(X)=H(X)-X$$

此时原最优解映射被表示为

$$H(X)=F(X)+X$$

深度残差网络自动提取特征时，只要 $F(X)=0$，即完成了一个恒等映射 $H(X)=X$，显然，使网络去拟合确定的函数 $F(X)=0$ 比逼近某个函数 $H(X)$ 容易。

在深度残差网络中，$F(X)+X$ 通过快捷连接来实现，结构如图 2(a)所示。快捷连接跳过 2 或 3 个卷积层，自身映射到叠加层相加作为输出。文献[16]进一步提出了预激活机制的深度残差网络 Resnetv2，在卷积操作前进行批规范化和激活操作，如图 2(b)所示，采用预激活方式使深度残差网络加强对模型的正则化，并加快网络收敛。

2　针对图像特征的行人检测方法

2.1　基于图像的行人特征分析

人类通过双眼观察事物，视像设备记录人类眼中的事物，这些事物都是“真实”的，这是针对行人图像特征的行人检测方法设计的灵感来源。计算机视觉是机器模拟生物视觉的方法，如果给机器更贴近原生的输入，则更有利于机器“理解”世界。

通过对行人图像进行分析，如图 3 所示，可归纳以下特点：

a）行人个体的横向(X)特征表达比纵向(Y)特征表达少，如图 3(d)所示。行人身高和身宽比约为 3:1，横向特征表达较少，增加横向特征表达有助于行人识别。

b）行人群体在没有标注的图像上无法体现深度信息，因此如图 3(a)人群可以看做行人特征的横向堆叠，增加图像宽度，有利于行人群体特表达。



图 3　卷积神经网络的图像预处理比例

另外，视像设备多采用 16:9 长宽比例采集数据，视像数据多为图 3(a)(d)形式，特别是图 3(a)画幅是视频采集常用画幅。而目标检测方法通常输入处理后的正方形图像，这种处理加大了行人横向和纵向特征差异，从图 3(b)(c)(e)(f)可知横幅图像的失真比竖幅图像的失真程度更高。Overett 等人在 Haar[17]特征行人识别研究中也有类似工作，认为对于 Haar 特征方法增加样本宽度的性能好于增加样本高度的性能。综上，本文采用了改变卷积神经网络输入的长宽比例，利用矩形输入网络提取更丰富的横向特征，经实验验证，本文选择了 288×608 分辨率图像作为输入，这个分辨率的图像与 416×416 图像有相近像素数，不影响实时性。

2.2　基于维度聚类的预测框选取

在基于区域提名网络的目标检测方法中，先验预测框的设置往往是手工设计而成，虽然网络在迭代过程中也能拟合随机性较大，使网络拟合缓慢，而且更容易使模型收敛到局部最优，文献[13]针对这个问题提出了维度聚类的方法选取先验预测框。本文采用了 K-means 聚类方法对数据集进行分析，确定先验框个数和大小。

在实现 K-means 聚类时，没有采用闵可夫斯基距离变式作为评价指标，而是通过交并比衡量，这样可以避免框的尺寸对 K-means 损失的影响。并且计算交并比时没有采用直接的框的的长宽计算，而是通过单位网格将其标准化。最终，相异度计算公式为

$$d(box,centrd)=1-IoU(box,centrd)$$

其中：$centrd$ 表示簇中心，box 表示样本。$IoU(box,centrd)$ 表示簇中心框和聚类框的交并比。交并比 IoU 表示预测框的准确程度，公式表示为

$$IoU(bb_{gt},bb_{dt})=\frac{bb_{gt}\cap bb_{dt}}{bb_{gt}\cup bb_{dt}}$$

其中：bb_{gt} 表示真实框，bb_{dt} 表示预测框。

本文通过实验对比了 K=[2~9]八组预测框对模型定位能力的影响，见图 4，均衡网络性能和检测实时性，最终选择预测框个数为 5 个。

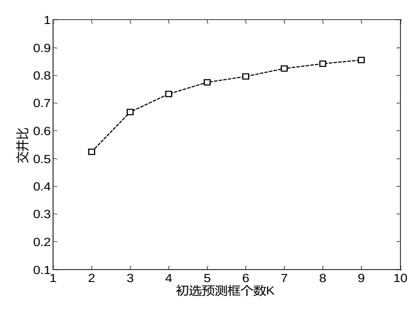


图 4　预测框个数与交并比关系

2.3　Res-P-YOLO 网络设计

本文基于 YOLOv2 目标检测模型，把 Resnet v2-50 分类网络与 YOLOv2 的检测方法结合。首先去掉 50 层预激活深度残差网络最后的一个平均池化层和一个全链接层，此时最后的卷积层输出维度为 2048，为了避免高维度带来的计算量，在此后串联一个 1×1×1024 卷积层,用来降低维度以减少运算量；其次有针对性的将深度残差网络的输入设计为 288×608，保留图像更多横向特征；然后接入 YOLOv2 的检测网络，并将分类网络中 18×38×512 特征图作为细粒度特征与检测网络的特征图重组，提高网络小物体识别能力；最后根据识别目标数目和预测框设置调整模型，完成 Res-P-YOLO 行人识别方法的设计，结构图如图 5 所示。

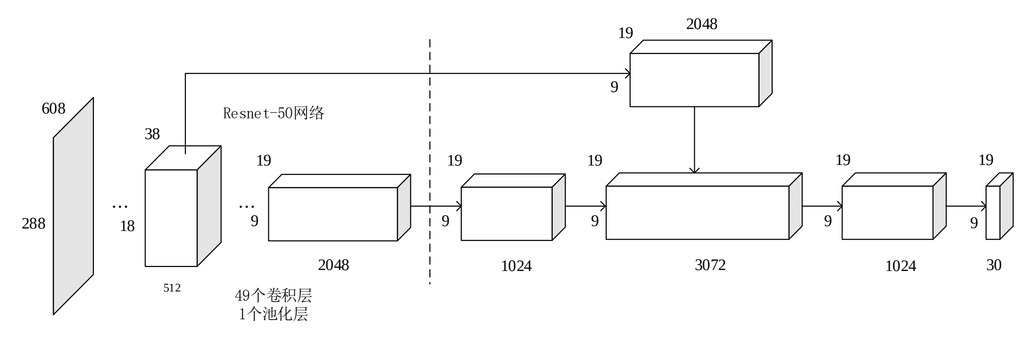


图 5　Res-P-YOLO 网络

在训练过程中，Res-P-YOLO 端到端回归行人位置坐标的同时，也进行了预测框中行人的置信度 $Conf(Person)$ 计算，置信度包含了预测框的准确程度和框内存在行人的可能性 $\Pr(Person)$ 为

$$Conf(Person) = \Pr(Person) \times IoU_{dt}^{gt}$$

因此，在训练过程中 Res-P-YOLO 在计算损失时需要衡量行人位置、预测框准确度和行人可能性三方面，故模型的损失函数表示为

$$\begin{aligned} loss(person) = {} & \lambda_{coord} \sum_{i=0}^{W\times H} \sum_{j=0}^{B} \Delta_{ij}^{pers} ((x_i - \hat{x}_i)^2 \\ & + (y_i - \hat{y}_i)^2 + (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2) \\ & + \sum_{i=0}^{W\times H} \sum_{j=0}^{B} \Delta_{ij}^{pers} (c_i - \hat{c}_i)^2 \\ & + \lambda_{nopers} \sum_{i=0}^{W\times H} \sum_{j=0}^{B} \Delta_{ij}^{nopers} (c_i - \hat{c}_i)^2 \\ & + \sum_{i=0}^{W\times H} \Delta_i^{pers} (p_i - \hat{p}_i) \end{aligned}$$

其中：W,H 表示网络最终特征图网格尺寸，B 表示每个网格的预测框个数，x,y,w,h 表示框的中心和宽高；C_i 表示预测框定位到行人的置信度，$\hat{C}_i$ 表示框内真实存在行人的置信度，p_i 表示预测行人置信度，$\hat{p}_i$ 真实存在行人置信度。$\lambda_{coord}, \lambda_{nopers}$ 分别表示位置预测和行人预测正则化惩罚系数，Δ_{ij}^{pers} 表示第 i 个网格的第 j 个预测框存在行人的可能性，Δ_i^{pers} 表示判定第 i 个网格存在行人的概率。

3　方法验证与实验对比

3.1　实验环境

本文的实验环境如下，CPU：Intel Core i7-6800K；内存：16G DDR4；GPU：Nvidia Geforce GTX1070；操作系统：64 位 Ubuntu16.04 LTS；实验框架为 Darknet 开源框架。数据集选用了 INRIA 行人数据集和提取自 PASCAL VOC 数据集的行人数据。在该实验环境下，Res-P-YOLO 在 1080p 的 mp4 视频上的检测速度可达 31.2 帧/秒。

3.2　行人检测模型评价方法

本文使用每张图片误检率(false positive per image，FPPI)和漏检率(Miss Rate，MR)指标评价行人检测方法，使用 ROC 曲线直观表示。该指标通过整张图像评估反映整个行人检测系统的性能，FPPI-MR 指标越低表明行人检测方法性能越好。漏检率和每张图片误检率计算方法分别如下

$$MR = \frac{FN}{FN + TP}$$

$$FPPI = \frac{FP}{N_{test_data}}$$

其中：FN 表示实际为行人，被判定非行人的个数；TP 表示实际为行人，被判定行人的个数；FP 表示实际非行人，被判定行人的个数；N_{test_data} 表示测试集图片张数。

3.3　混合数据集训练效果验证

INRIA 行人数据集是目前被应用最多的公开数据集，训练集标注准确，场景丰富，分为测试集和训练集两部分。训练集包含行人样本 2416 个，分布在 614 张图片上；测试集包含行人样本 1126 个，分布在 288 张图片上。PASCAL VOC 数据集是知名的 20 分类的公开多目标数据集，本文根据标注信息，从混合的数据集中手工提取出 8102 张行人图片以及对应标注信息。

深度学习方法需要大量数据训练，少量数据容易造成过拟合，虽然在该数据集上有较好表现，但实际应用时泛化能力较差。针对这个问题，本文提取了 PASCAL VOC 数据集的全部行人数据用来辅助训练，提高本方法的泛化能力。本文使用 INRIA 和 VOC 混合训练模型与仅使用 INRIA 训练模型进行实验对比，如图 6 所示，实验表明使用混合训练的方法能使行人检测的每张图片漏检率降低 5%。因此，本文综合了 INRIA 行人数据集和 PASCAL VOC 数据集行人图片部分，共有样本图片 8716 张，本文选取 PASCAL VOC 训练集全部行人图片和 INRIA 训练集作为实验的训练集；选取 PASCAL VOC 测试集全部行人图片作为实验的验证集；采用 INRIA 测试集作为实验的测试集。

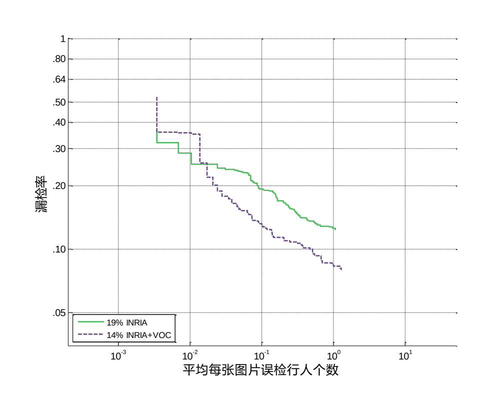


图 6　混合数据集训练和单一数据集训练模型性能对比

3.4　矩形输入性能对比

在设计矩形输入时有两种思路：一是在原始输入 416×416 基础上进行横向扩展，但会导致分辨率骤增，无法验证矩形输入对网络的影响，并会影响实时性；二是自行设计新的矩形尺寸使像素数尽量与 416×416 相近，在保证实时性的同时可以验证矩形输入对网络的影响。

本文验证采用方法二，矩形输入的设计见表 1。

表 1　网络矩形输入设计

分辨率	长宽比	像素数
416×416	13:13	173056
352×480	11:15	168960
352×544	11:17	191488
288×544	9:17	156672
288×608	9:19	175104

显然 288×608 像素数与 416×416 最为相近，故采用 288×608 和 416×416 两种输入尺寸进行实验比对。为了加快验证

速度，采用了 yolov2 精简方法 tiny-yolo[18]在上述训练集上进行行人检测定量分析实验，实验均采用默认设置，仅修改输入分辨率设置，从图 7 可以看出，采用 288×608 输入的行人检测方法有明显提升，召回率提高 6.37%，平均交并比提升 2.22%。

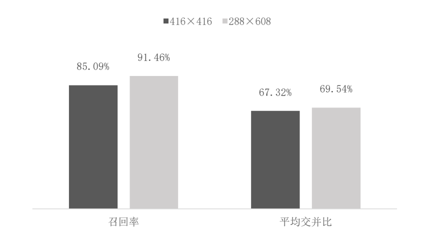


图 7　不同输入尺寸对网络性能的影响

3.5　实验结果与对比

Res-P-YOLO 网络的构建和实现基于开源的深度学习框架 Darknet[18]。训练时学习率采用均匀分布策略(polynomial decay)，初始学习率为 0.001，在 40000 和 60000 次迭代后分别在乘以 0.1,动量系数为 0.9，权值衰减系数为 0.0005，并采用了随机调整曝光、饱和度、色调数据增强的方法。实验以 64 张图片/批进行迭代，迭代 80000 次，如图 8 所示，经验证集确认第 73800 次迭代结果作为最终权重模型。

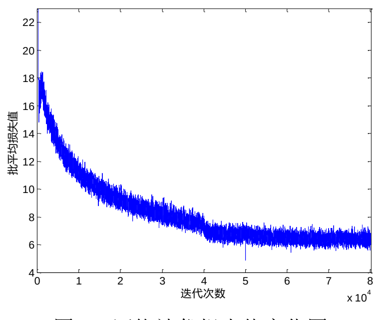


图 8　网络迭代损失值变化图

为了综合评价本文算法有效性，本节对综合上述工作，在 INRIA 行人测试集上对本文方法与主流的 LatSVM[19]、ACF[20]、RandForest[21]、ConvNet[22]等行人检测方法以及 YOLOv2 方法的检测性能指标进行评估，得出 FPPI-MR 关系的 ROC 曲线，ROC 曲线下的面积越小，说明行人检测方法检测性能越好。实验结果如图 9 和表 2 所示。

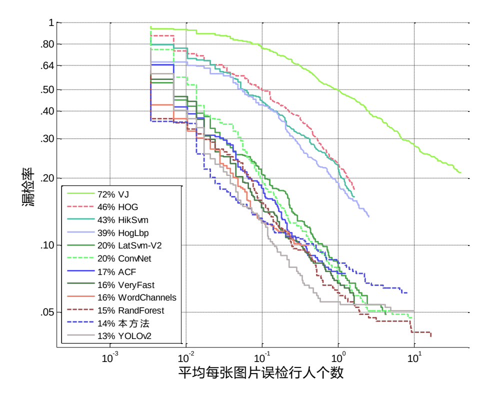


图 9　本算法和主流行人算法的 ROC 曲线图

从图 9 本方法和 YOLOv2 的 ROC 曲线可以看出，本方法与 YOLOv2 相比，当 FPPI<10^{-1}时本方法明显优于 YOLOv2 方法。从曲线总体来看，本方法的 FPPI 指标较 YOLOv2 低约 1%，因为 YOLOv2 采用了 ImageNet 预训练模型训练，而本方法自行设计的网络没有相应预训练模型，ImageNet 预训练模型对调参影响巨大，因此 1%的差距在理论上是合理的。

从表 2 可以看出，本方法在 INRIA 数据集上取得了较好的性能表现，每张图片误检率仅 13.86%，不同程度上优于图表中其他传统算法，特别是在 FPPI<10^{-1}时，本算法的行人检测性能明显优于其他算法。其原因在于大多数手工设计特征的行人检测方法对行人的表征能力较差，而本算法针对图片行人特征设计了有助于图像中行人特征表达的网络输入和能够更好提取行人图像特征的深度残差网络，并通过混合数据集、聚类初选预测框方法加强训练，使本方法能够自主学习到更有效的行人特征，进而进一步降低每张图片误检率。

表 2　主流行人算法检测率对比

行人检测方法	FPPI(%)
VJ	72.4810
HOG	45.9788
HikSvm	42.8188
HogLbp	39.0968
LatSvm-V2	19.9641
VeryFast	15.9591
ConvNet	19.8934
ACF	17.2809
RandForest	15.3714
WordChannels	15.9511
本方法	13.8589

4　结束语

本文针对行人自身形状特征和在图像中的分布特征提出了一种基于深度残差网络的行人检测方法。针对行人形状和分布特点，提出了一种矩形输入的卷积神经网络分类模型，加强了行人特征表达；为了提高模型特征提取能力，提出了一种基于预激活深度残差网络的行人检测方法。在方法设计时，为了提高模型准确率和泛化能力，提出了混合行人数据集训练的方式降低每张图片误检率，并且通过聚类分析预测框的方法重新为模型设计了初始预测框，提高行人定位能力并加快收敛。实验结果表明，本算法拥有良好的实时性和泛化能力的同时，在公开的行人数据集下本方法较主流行人检测方法的准确率有显著提升。本文仅是对基于卷积神经网络的行人检测方法的初步探索，如何在保证实时性的进一步提高准确率和泛化能力是下一步探索的方向。

参考文献：

[1] 苏松志，李绍滋，陈淑媛，等. 行人检测技术综述 [J]. 电子学报, 2012, 40 (4): 814-820.

[2] Benenson R, Omran M, Hosang J, *et al.* Ten years of pedestrian detection, what have we learned? [C]// Proc of European Conference on Computer Vision. Springer, 2014: 613-627.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005: 886-893.

[4] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns [J] , IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24 (7): 971-987.

[5] 张慧，王坤峰，王飞跃. 深度学习在目标视觉检测中的应用进展与展望 [J]. 自动化学报, 2017, 43 (8): 1289-1305.

[6] 周飞燕，金林鹏，董军. 卷积神经网络研究综述 [J]. 计算机学报, 2017, 40 (6): 1229-1251.

[7] Girshick R, Donahue J, Darrell T, *et al.* Region-Based Convolutional Networks for Accurate Object Detection and Segmentation [J]. IEEE Transs on Pattern Analysis & Machine Intelligence, 2016, 38 (1): 142.

[8] Girshick R. Fast R-CNN [C]// Proc of IEEE International Conference on Computer Vision. 2015: 1440-1448.

[9] Ren S Q, He K M, Girshick R, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 39 (6): 1137−1149.

[10] He K, Zhang X, Ren S, *et al.* Spatial pyramid pooling in deep convolutional networks for visual recognition. [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2015, 37 (9): 1904-16.

[11] Dai J, Li Y, He K, *et al.* R-FCN: object detection via region-based fully convolutional networks [C]// Neural Information Processing Systems. 2016: 379-387.

[12] Redmon J, Divvala S, Girshick R, *et al.* You only look once: unified, real-time object detection [C]// Computer Vision and Pattern Recognition. 2016: 779-788.

[13] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [J]. arXiv preprint arXiv: 1612. 08242, 2016.

[14] Liu W, Anguelov D, Erhan D, *et al.* SSD: single shot multibox detector [C]// Proc of European Conference on Computer Vision. Springer, 2016: 21-37.

[15] He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition [C]// Computer Vision and Pattern Recognition. 2016: 770-778.

[16] He K, Zhang X, Ren S, *et al.* Identity mappings in deep residual networks [C]// Proc of European Conference on Computer Vision. Springer International Publishing, 2016: 630-645.

[17] Overett G, Petersson L, Brewer N, *et al.* A new pedestrian dataset for supervised learning [C]// Proc of Intelligent Vehicles Symposium. 2011: 373-378.

[18] Redmon J. Darknet: Open Source Neural Networks in C [DB/OL]. [2017-10-25]. http://pjreddie. com/darknet/.

[19] Felzenszwalb P, McAllester D, Ramanan D. A Discriminatively Trained, Multiscale, Deformable Part Model [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1-8.

[20] Dollar P, Appel R, Belongie S, *et al.* Fast Feature Pyramids for Object Detection [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2014, 36 (8): 1532.

[21] Marin J, Vazquez D, Lopez A M, *et al.* Random Forests of Local Experts for Pedestrian Detection [C]// Proc of IEEE International Conference on Computer Vision. 2014: 2592-2599.

[22] Sermanet P, Kavukcuoglu K, Chintala S, *et al.* Pedestrian detection with unsupervised multi-stage feature learning [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2013: 3626-3633.